



Intro to Data Science

- Day 1 → 5th Sep (13:00-17:00):
An Overview of Data Science Technologies
- Day 2 → 12th Sep (10:00-15:00):
Modern Applications of Data Science Methods



Intro to Data Science

- Day 1 → 5th Sep (13:00-17:00):
An Overview of Data Science Technologies
- Day 2 → 12th Sep (10:00-15:00):
Modern Applications of Data Science Methods

Trainer: Pedro V Hernández Serrano
< Data Scientist - Maastricht University >

Format

- 1 Hour lecture (incl. Q&A)
- 20 mins. Set up
- 40 mins. Live examples (Part 1)
- 20 mins. Questionnaire + Break
- 40 mins. Live examples (Part 2)
- 1 Hour hands-on

An Overview of Data Science Technologies

Pedro V Hernández Serrano
Workshop (Day 1) - Academy @Stamicarbon
05/09/2022

Trainer

Data Scientist

- Certified Cloud Solutions Engineer (GCP)
- Certified Data Protection Officer (ECPC)
- MSc Data Science, BSc Actuarial Science

Current

- Faculty **Data Steward** at Maastricht University
- **Cloud Architect** at European Trade Union Institute (ETUI)
- **Research Fellow** at Netherlands eScience Center
- **Freelance** Data Science & Data Governance Bootcamps



Contact:

Site: pedrohserrano.com

LinkedIn: linkedin.com/in/pedrohserrano/

Twitter: [@pedrohserrano](https://twitter.com/@pedrohserrano)

Lecture

Day 1

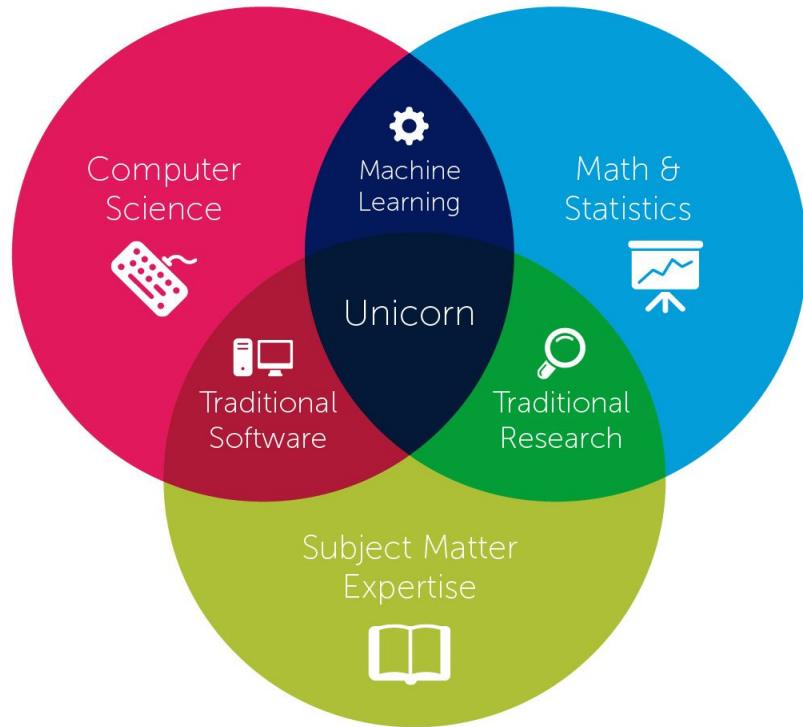
- What is Data Science
 - Data Science:
Technologies vs Methods
 - Applications
 - EDA
 - Cloud Services
 - Data Annotation
 - Knowledge Databases
 - Data Science with Python
-

Goals

- Historical recap and context around data science
- Explore some main technologies used in data science
- Introduce the basics of the python programming environment.
- Develop a general understanding of data formats and representations
- Get an overview of some python visualization packages.
- Know limitations and caveats of available interactive python libraries
- We will use open source analytics tools
- We will learn through doing.

What is Data Science?

What is Data Science



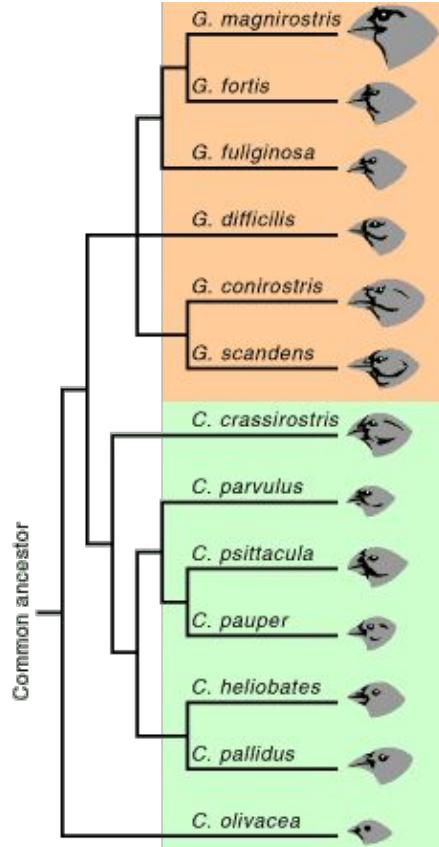
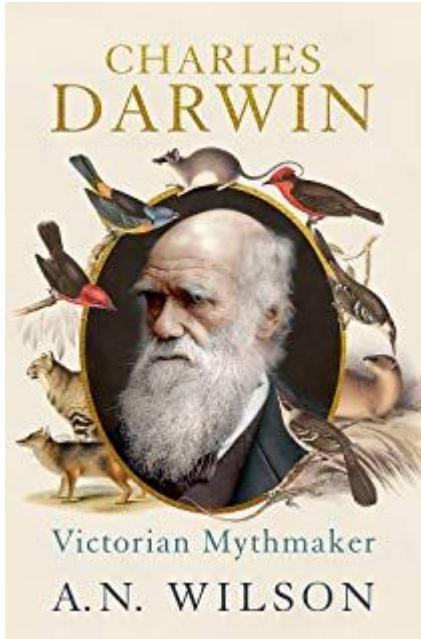
Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.



Is Data Science a new thing?



Observation and categorization



The reality is that we stored information for analytical purposes for quite a while
Many people in academia say that every scientist is a data scientist.

Statistics early stage

The Years of our Lord	The Table of CASUALTIES.																	
	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1661	1662	1663	1664
live and Still-born	335	329	227	254	389	381	584	433	453	419	463	407	421	544	499	439	410	445
and Fever	910	835	829	280	834	861	974	743	892	849	1170	909	1095	579	710	661	671	704
die and Suddenly	65	74	64	74	100	111	115	86	91	102	113	91	67	23	36	17	24	20
die	4	1	1	3	7	2	4	5	9	3	8	13	8	10	15	6	4	4
die	3	2	5	1	3	4	3	2	7	3	6	7	2	5	3	4	1	2
die Flux, Scouring and Flux	155	170	801	189	833	702	200	386	165	368	362	332	251	449	436	352	345	273
die Scalded	2	0	10	5	11	6	3	7	10	5	7	4	6	3	10	7	5	1
die	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1
die Gingivitis and Fistula	28	29	31	25	31	36	37	73	31	24	35	63	51	20	16	23	27	30
die	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
die Sore-mouth and Thrush	66	28	51	47	68	51	73	72	44	81	19	27	73	69	6	4	4	1
die	161	106	116	117	200	213	239	192	177	201	236	229	225	194	150	152	112	171
die and Infants	1369	1254	1065	930	1137	1250	1050	1343	1069	1182	1146	838	1121	1097	1374	1035	2155	2130
die Wind	103	71	85	82	74	101	80	101	85	120	123	173	116	107	48	57	37	50
die Cough	2473	2100	2185	1985	2350	2410	2250	2553	2000	1842	1972	1610	1902	3424	1827	1910	1717	1797
die	634	491	530	493	569	653	660	523	703	1027	807	841	743	1031	52	87	18	84
die	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
die Stone	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
die and Tympany	185	434	424	505	444	356	617	704	660	706	633	931	646	572	235	252	270	260
die	47	40	30	37	43	50	57	30	43	49	63	60	57	48	43	33	29	34
die and drinking	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
die scared	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
die and in Bath	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
die Suckus	3	2	2	3	3	4	1	4	3	1	4	5	3	10	7	2	3	0
die and small Pox	159	400	1150	1124	1255	1279	119	912	1194	813	835	402	1523	354	72	40	58	534
die dead in the Streets	6	6	6	6	6	7	9	16	4	3	4	5	11	3	16	33	30	0
die Pox	15	25	15	21	21	20	20	20	20	23	23	53	51	21	17	12	12	7
die ghred	4	4	1	1	3	1	2	1	1	1	1	1	1	1	1	1	1	1
die	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
die and made-away themselves	11	10	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
die Ach	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
die	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

18th century:

Systematic collection of demographic and economic data by states.

For tax or military use...



So for decades statisticians,
economists, demographers, were
very happy doing data analysis, till...

The PC happened

No manual analysis
anymore

thedayintech.wordpress.com/2013/03/08/4995-00-for-10-megabytes/



**The IBM Personal Computer XT.
More power to the person.**

Plenty of muscle. That's what the new IBM Personal Computer XT means to a person with heavyweight data to manage. Because one of the XT's many strong points is a 10-million-character fixed disk drive that helps give you the power to pump more productivity into your business. What's so special about a fixed disk? Exactly that. It's *already* fixed inside the system, with the capacity to store the facts, figures, names and numbers you need to work with.

(Rather than go from diskette to diskette, store up to 5,000 pages of text or up to 100,000 names and addresses in one place.)

Yet there's more built into the XT than its fixed disk. Reliability and quality are built in as well. Plus more than 30 years of IBM experience.

A new level of price/performance. And a remarkable compatibility of both software and hardware with the original IBM Personal Computer. So, with the introduction of XT comes a special tool designed to help you be more productive in high-volume applications.

WHAT'S THE DIFFERENCE?



BASE SYSTEM™

User Memory
64KB (expandable to 640KB)

Auxiliary Memory
Up to two 5½" 300KB/500KB or 1.2MB/1.4MB diskette drives
optional



BASE SYSTEM™

User Memory
128KB (expandable to 640KB)

Auxiliary Memory
One 35 megabyte hard disk
or up to two 5½" 300KB/500KB
diskette drives standard

* An expansion unit can also be added to both 16-bit (8088) systems for 6 more expansion slots. Added to the IBM Personal Computer, it can handle two 12 megabyte hard disk drives or up to four 5½" 300KB diskette drives. XT is the same base model without fixed disk drive or a total of 20 megabytes.

Another tool for modern times to
keep you going strong.

To find out where you can see the
IBM Personal Computers, call 800-447-4700.
In Alaska or Hawaii, 800-447-0890.





Then new scientists, chemists,
econometricians were very happy
doing data analysis, till...

The Big Data happened

We needed to study
the technologies to
harness data



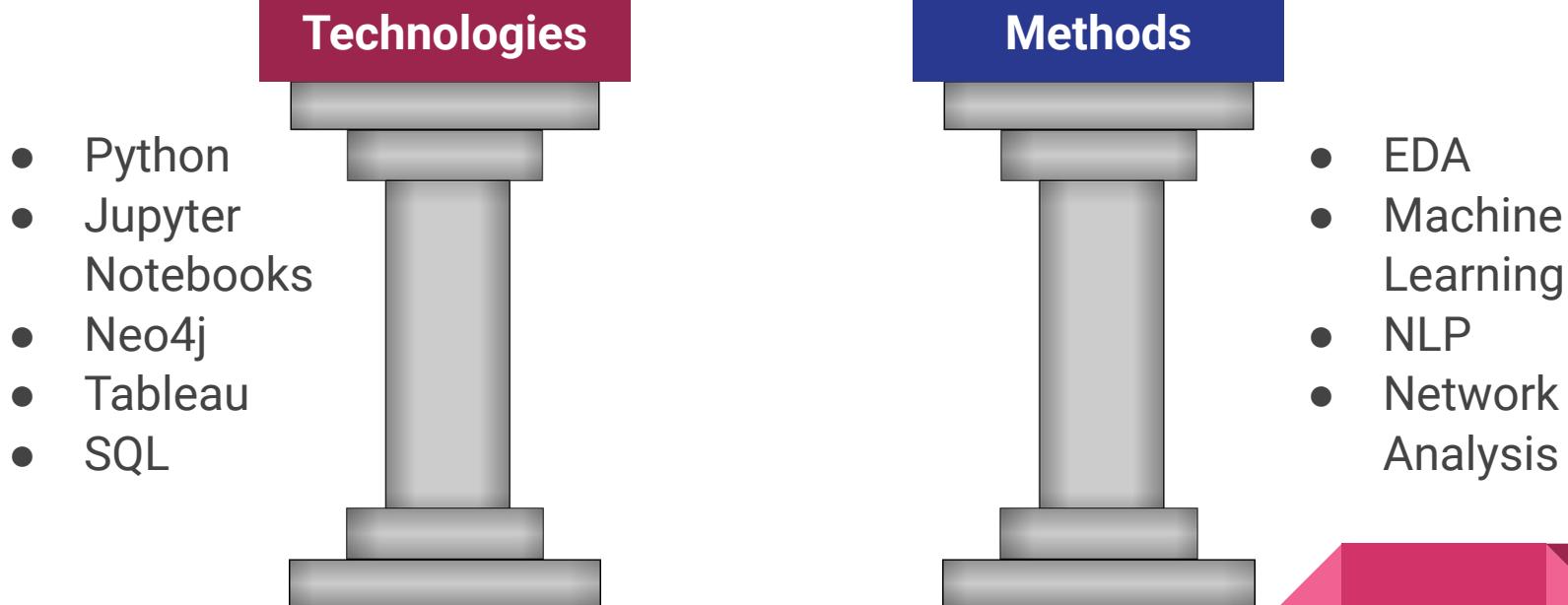
The Data Science profession was born



The Data Scientist aimed to
use “new” **technologies**
to adapt “old” **methods**

Data Science: Technologies vs Methods

Data Science pillars



Technologies

Technologies

Manual

	Couriers	Lettres	Envois	Parcels	Minéraux	Grains	Graine	Graine	Graine
1	100	100	100	100	100	100	100	100	100
2	100	100	100	100	100	100	100	100	100
3	100	100	100	100	100	100	100	100	100
4	100	100	100	100	100	100	100	100	100
5	100	100	100	100	100	100	100	100	100
6	100	100	100	100	100	100	100	100	100
7	100	100	100	100	100	100	100	100	100
8	100	100	100	100	100	100	100	100	100
9	100	100	100	100	100	100	100	100	100
10	100	100	100	100	100	100	100	100	100
11	100	100	100	100	100	100	100	100	100
12	100	100	100	100	100	100	100	100	100
13	100	100	100	100	100	100	100	100	100
14	100	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100	100
16	100	100	100	100	100	100	100	100	100
17	100	100	100	100	100	100	100	100	100
18	100	100	100	100	100	100	100	100	100
19	100	100	100	100	100	100	100	100	100
20	100	100	100	100	100	100	100	100	100
21	100	100	100	100	100	100	100	100	100
22	100	100	100	100	100	100	100	100	100
23	100	100	100	100	100	100	100	100	100
24	100	100	100	100	100	100	100	100	100
25	100	100	100	100	100	100	100	100	100
26	100	100	100	100	100	100	100	100	100
27	100	100	100	100	100	100	100	100	100
28	100	100	100	100	100	100	100	100	100
29	100	100	100	100	100	100	100	100	100
30	100	100	100	100	100	100	100	100	100
31	100	100	100	100	100	100	100	100	100
32	100	100	100	100	100	100	100	100	100
33	100	100	100	100	100	100	100	100	100
34	100	100	100	100	100	100	100	100	100
35	100	100	100	100	100	100	100	100	100
36	100	100	100	100	100	100	100	100	100
37	100	100	100	100	100	100	100	100	100
38	100	100	100	100	100	100	100	100	100
39	100	100	100	100	100	100	100	100	100
40	100	100	100	100	100	100	100	100	100
41	100	100	100	100	100	100	100	100	100
42	100	100	100	100	100	100	100	100	100
43	100	100	100	100	100	100	100	100	100
44	100	100	100	100	100	100	100	100	100
45	100	100	100	100	100	100	100	100	100
46	100	100	100	100	100	100	100	100	100
47	100	100	100	100	100	100	100	100	100
48	100	100	100	100	100	100	100	100	100
49	100	100	100	100	100	100	100	100	100
50	100	100	100	100	100	100	100	100	100
51	100	100	100	100	100	100	100	100	100
52	100	100	100	100	100	100	100	100	100
53	100	100	100	100	100	100	100	100	100
54	100	100	100	100	100	100	100	100	100
55	100	100	100	100	100	100	100	100	100
56	100	100	100	100	100	100	100	100	100
57	100	100	100	100	100	100	100	100	100
58	100	100	100	100	100	100	100	100	100
59	100	100	100	100	100	100	100	100	100
60	100	100	100	100	100	100	100	100	100
61	100	100	100	100	100	100	100	100	100
62	100	100	100	100	100	100	100	100	100
63	100	100	100	100	100	100	100	100	100
64	100	100	100	100	100	100	100	100	100
65	100	100	100	100	100	100	100	100	100
66	100	100	100	100	100	100	100	100	100
67	100	100	100	100	100	100	100	100	100
68	100	100	100	100	100	100	100	100	100
69	100	100	100	100	100	100	100	100	100
70	100	100	100	100	100	100	100	100	100
71	100	100	100	100	100	100	100	100	100
72	100	100	100	100	100	100	100	100	100
73	100	100	100	100	100	100	100	100	100
74	100	100	100	100	100	100	100	100	100
75	100	100	100	100	100	100	100	100	100
76	100	100	100	100	100	100	100	100	100
77	100	100	100	100	100	100	100	100	100
78	100	100	100	100	100	100	100	100	100
79	100	100	100	100	100	100	100	100	100
80	100	100	100	100	100	100	100	100	100
81	100	100	100	100	100	100	100	100	100
82	100	100	100	100	100	100	100	100	100
83	100	100	100	100	100	100	100	100	100
84	100	100	100	100	100	100	100	100	100
85	100	100	100	100	100	100	100	100	100
86	100	100	100	100	100	100	100	100	100
87	100	100	100	100	100	100	100	100	100
88	100	100	100	100	100	100	100	100	100
89	100	100	100	100	100	100	100	100	100
90	100	100	100	100	100	100	100	100	100
91	100	100	100	100	100	100	100	100	100
92	100	100	100	100	100	100	100	100	100
93	100	100	100	100	100	100	100	100	100
94	100	100	100	100	100	100	100	100	100
95	100	100	100	100	100	100	100	100	100
96	100	100	100	100	100	100	100	100	100
97	100	100	100	100	100	100	100	100	100
98	100	100	100	100	100	100	100	100	100
99	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100

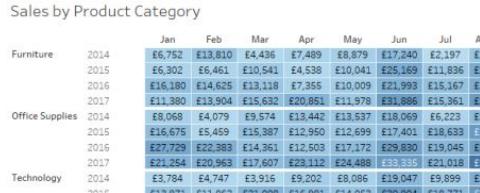
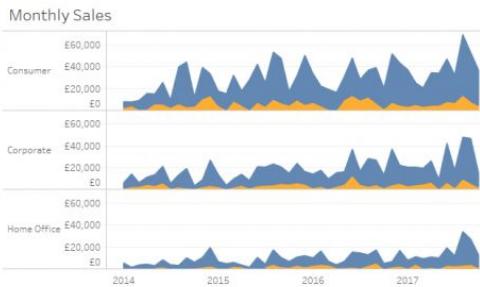
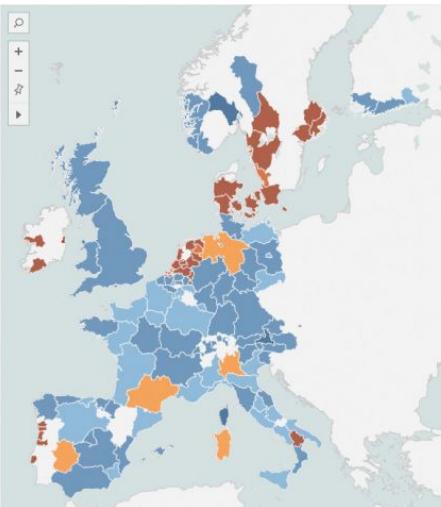


User oriented

Purpose oriented
Software
Programming Language



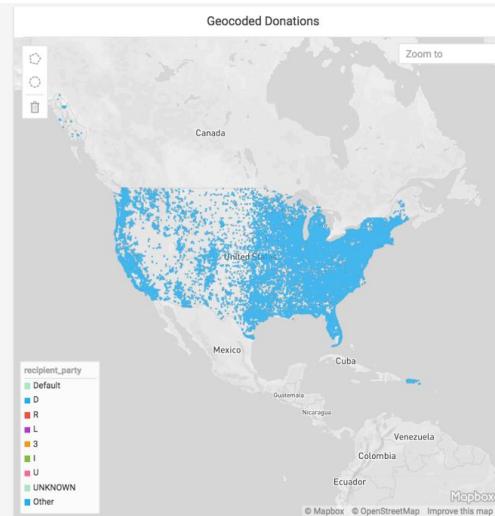
Purpose Oriented Technologies



Python

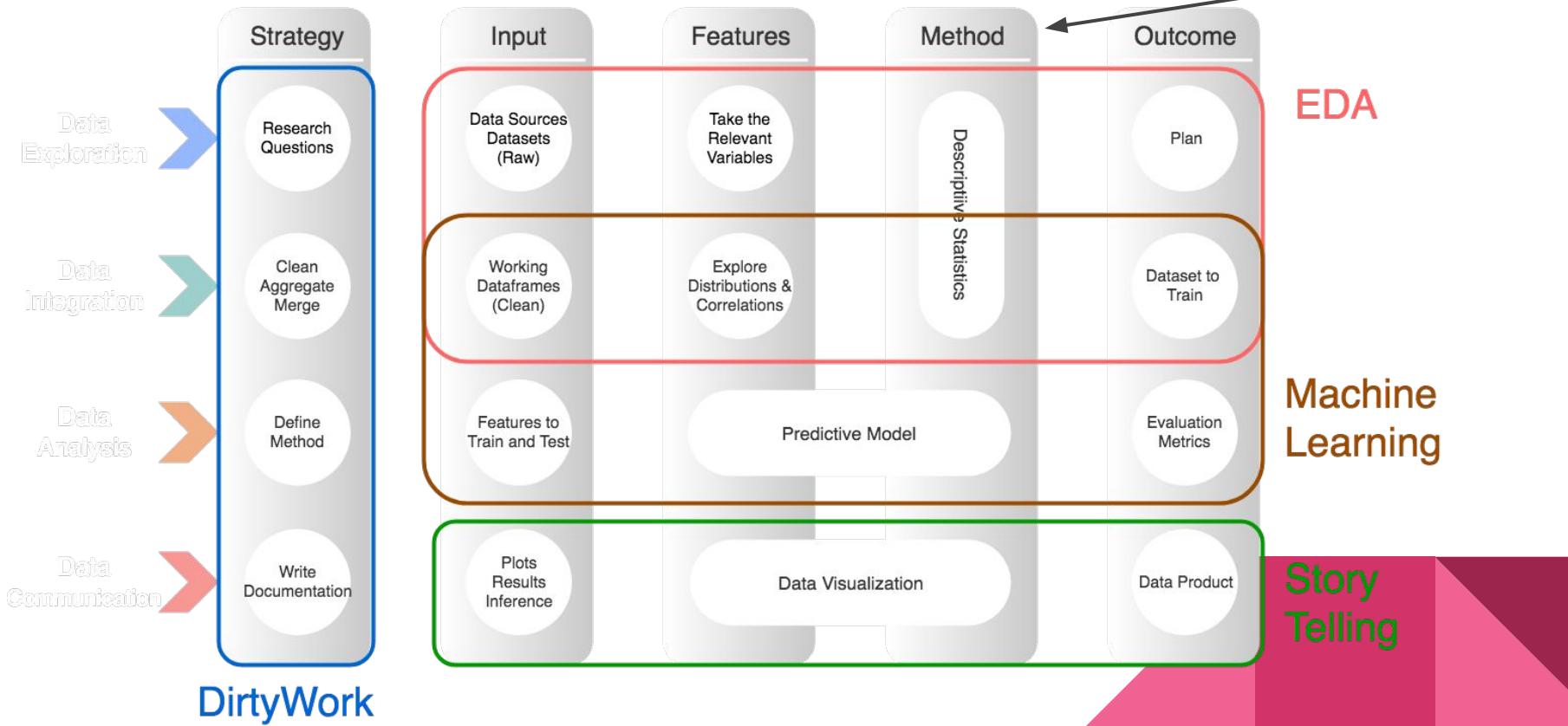


Tableau



Methods

Methods in Data Science



Nutshell

Technologies in this workshop

- Python + Notebooks
- Neo4j

Methods in this workshop

- EDA
- Machine Learning



Applications

Applications

Technologies + Methods

EDA

Simulations

Data Annotation

Knowledge Database

Exploratory Data Analysis (EDA)

python 3.6 | 3.7 | 3.8 | 3.9 license MIT DOI 10.5281/zenodo.5117896

Facebook Ads Data Analysis

Methodologies and limitations of using the Facebook Graph API to access the Facebook Ads library performing data analysis on the 2021 Dutch General Elections getting key insights of advertisement practices.

[See the full analysis here](#)

The image displays three examples of Facebook ads from the 2021 Dutch General Elections. Each ad is shown with its launch date, a thumbnail, a brief description, and a 'See ad details' button.

- Launched in March 2021:** Inactive ad for 'Woonbond'. It was sponsored by Woonbond and paid for by Woonbond. The ad encourages people to choose a party over the 'ellenlange' (long) list. It features a man in a dark hoodie. Key metrics: Amount spent (EUR): <€100, Potential reach: >1M people.
- Launched in February 2021:** Inactive ad for 'Jonge Socialisten in de PvdA'. Sponsored by Jonge Socialisten in de PvdA, it encourages young voters to support the PvdA's NextGen list. It features a group of diverse young people. Key metrics: Amount spent (EUR): <€100, Potential reach: >1M people.
- Launched in March 2021:** Inactive ad for 'Wopke Hoekstra'. Sponsored by CDA, it urges voters to choose the CDA over the 'enzaamheid' (loneliness). It features a man in a suit. Key metrics: Amount spent (EUR): <€100, Potential reach: 5K-10K people.

The image features the European Commission logo at the top right. Below it, the title of the study is displayed in large, bold, white font on a blue background.

Exploratory study on Information Technology (IT) and Artificial Intelligence (AI) Tools for Monitoring Online Markets for Consumer Policy Purposes

European Commission Tender
Final Report
(JUST/2018/CONS/PR/CO01/0123)

Simulation

Create a synthetic dataset of workers

```
import crowded.simulate as cs

#define your parameters
total_workers = 40
alpha = 28
beta = 2
#create task dataset
df_workers = cs.Workers(alpha, beta).create(total_workers)
```

Assign easily and fairly workers to tasks

```
import crowded.simulate as cs

#workers per task should always be smaller than the number of workers
wpt = 5
#create assignment
df_tw = cs.AssignTasks(df_tasks, df_workers, wpt).create()
```

Compute Bayes probability and predict worker answers

```
import crowded.method as cm

#workers per task should always be smaller than the number of workers
wpt = 5
#create assignment
df_tw = cs.AssignTasks(df_tasks, df_workers, wpt).create()
```

Compute Bayes probability and Predict answers of the workers

```
import crowded.method as cm

#define the parameters
x = df_tw['prob_task'] #vector of probabilities of tasks
y = df_tw['prob_worker'] #vector of probabilities of workers
z = df_tasks['true_answers'].unique() #vector of valid answers in the experiment
#compute probability
cp = cm.ComputeProbability(x, y, z)
```

RESEARCH-ARTICLE

CrowdED: Guideline for Optimal Crowdsourcing Experimental Design



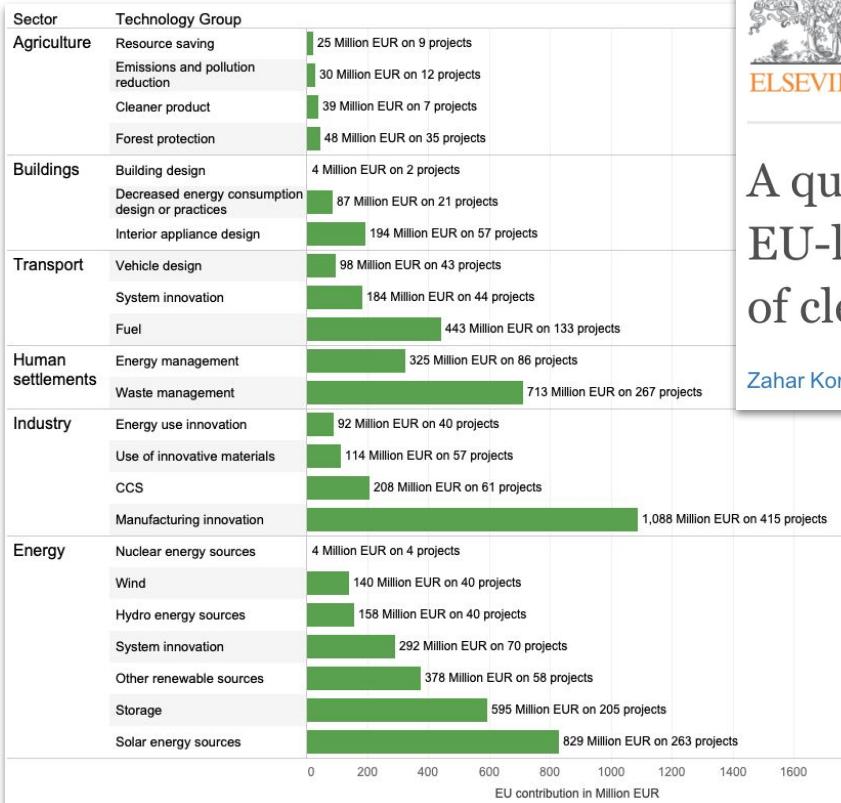
Authors: Amrapali Zaveri, Pedro Hernandez Serrano, Manisha Desai, Michel Dumontier

[Authors Info & Affiliations](#)

Publication: WWW '18: Companion Proceedings of the The Web Conference 2018 • April 2018 • Pages 1109–1116 • <https://doi.org/10.1145/3184558.3191543>

CrowdED: Guideline for Optimal Crowdsourcing Experimental Design
A Zaveri, Pedro Hernandez Serrano, M Desai, M Dumontier
Companion of the The Web Conference 2018 on The Web Conference 2018, 1109-1116

Data Annotation



Current Research in Environmental Sustainability

Volume 3, 2021, 100084

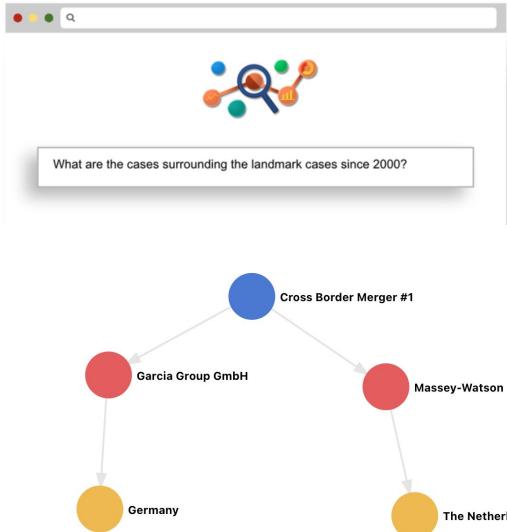


A qualitative-computational cataloguing of the EU-level public research and innovation portfolio of clean energy technologies (2014–2020)

Zahar Koretsky ^a , Pedro V. Hernández Serrano ^b , Seun Adekunle ^b , Michel Dumontier ^b



Knowledge Databases



The Case for a Linked Data Research Engine for Legal Scholars

Published online by Cambridge University Press: **04 November 2019**

Kody MOODLEY , Pedro V HERNANDEZ-SERRANO,
Amrapali J ZAVERI, Marcel GH SCHAPER, Michel DUMONTIER and
Gijs VAN DIJCK

<https://eu-corporate-mobility.org/>

Exploratory Data Analysis (EDA)



from Data to Viz

'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

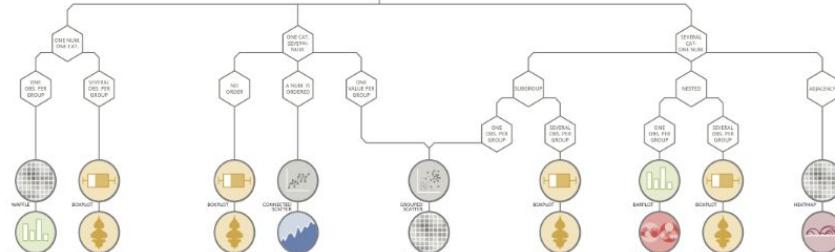
- 1 Identify what type of data you have.
- 2 Go to the corresponding decision tree and follow it down to a set of possible charts.
- 3 Choose the chart from the set that will suit your data and your needs best.

Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

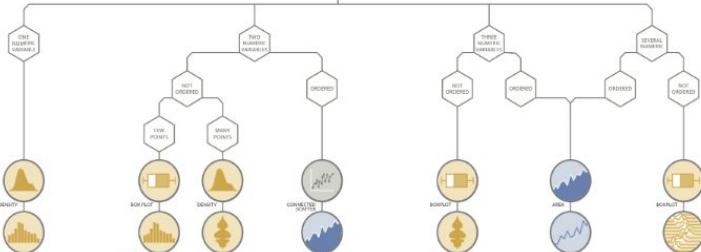
data-to-viz.com

data-to-viz.com/

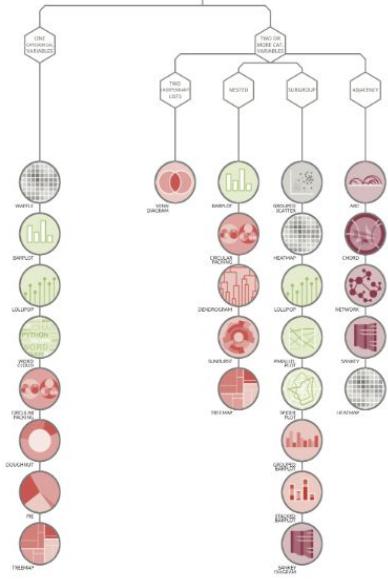
CATEGORIC AND NUMERIC



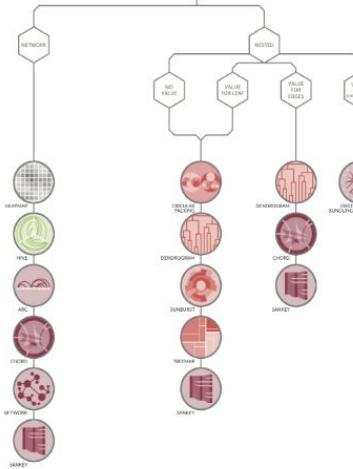
NUMERIC



CATEGORIC



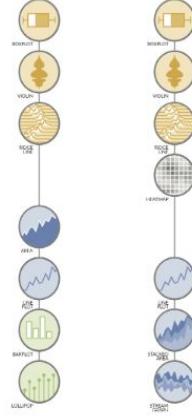
RELATIONAL



MAP



TIME SERIES



WHAT DO YOU WANT TO SHOW ?

- | | |
|--|---|
| ● Distribution | ● Evolution |
| ● Correlation | ● Maps |
| ● Ranking | ● Flow |
| ● Part of a whole | |

Technologies for EDA

Proprietary Software



- Shorter learning curve
- Automatic updates
- Specific formats
- Enterprise licence

Open Source



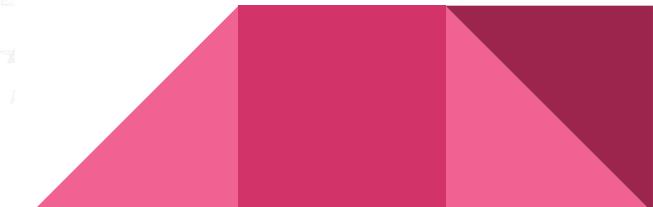
- Longer learning curve
- Harder Maintenance
- Interoperable
- Free

Cloud Services

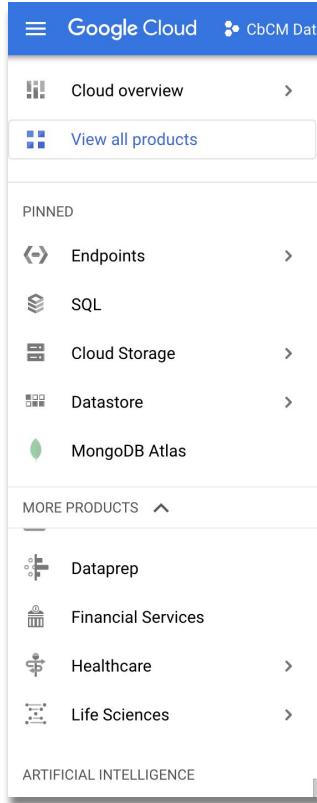
Cloud Services

- IaaS (Infrastructure as a service)
- Pay as you go business model
- Harmonized data centers aiming Net 0% CO₂

- Big data simulations should move out of your computer
- One-stop shop for all data problems
- Governance models & security



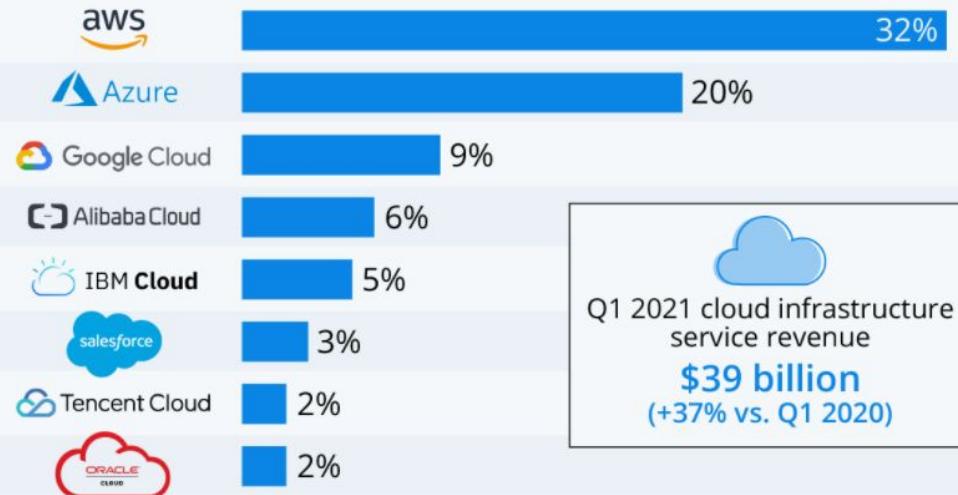
Cloud Services



All services in
one console

Amazon Leads \$150-Billion Cloud Market

Worldwide market share of leading cloud infrastructure service providers in Q1 2021*

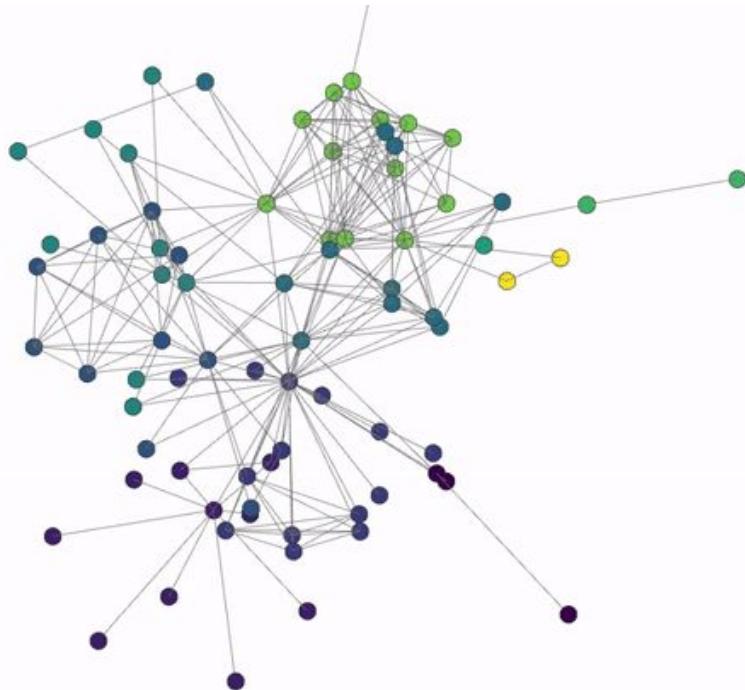


* includes platform as a service (PaaS) and infrastructure as a service (IaaS)
as well as hosted private cloud services

Source: Synergy Research Group

Understanding Data Annotation

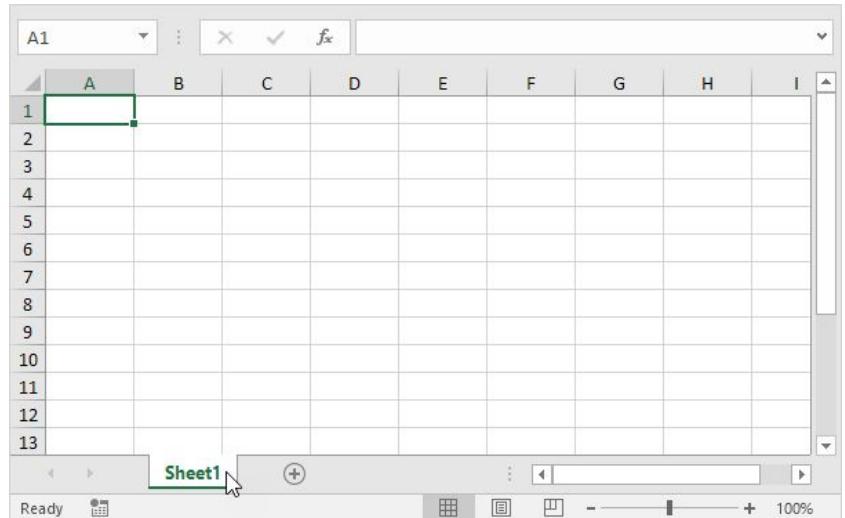
Data Stuff...



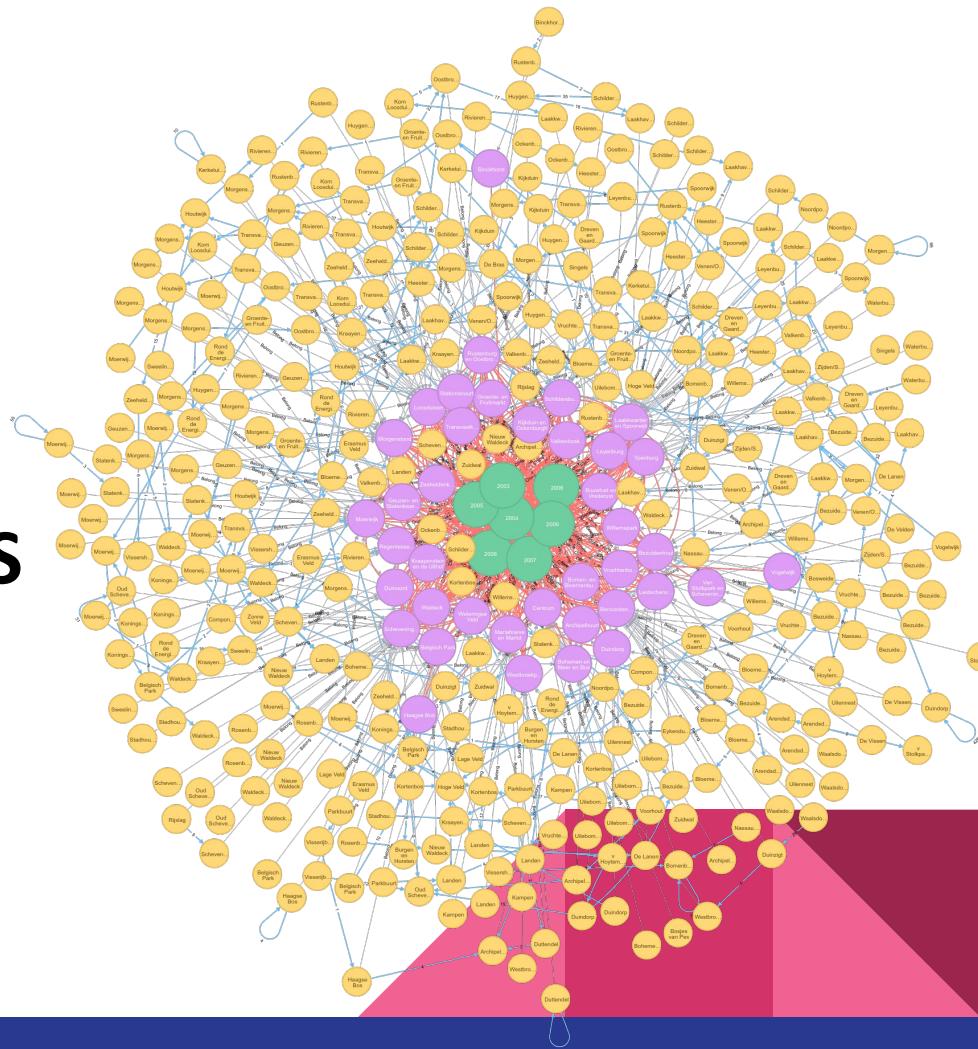
- Data Model
- Database
- Dataset
- Data Syntax
- **Data Format**

But before we can annotate and enhance data, we need to understand the terminology and especially the data formats

Data Format



VS



Tables

Columns

Country	Name	Year Passed	Executive	Category	Document
Japan	Act on the Improvem...	01/01/2015	Legislative	null	Energy De...
Macedoni...	Action Plan on Rene...	01/01/2015	Executive	null	Energy Sup...
Czech Rep...	Adaptation strategy ...	01/01/2015	Executive	Adapt...	Adaptation
Niue	Agriculture Sector Pl...	01/01/2015	Executive	null	Adaptation...
Thailand	Alternative Energy D...	01/01/2015	Executive	null	Energy Sup...
United Sta...	Clean Power Plan	01/01/2015	Executive	null	Energy Sup...
Malta	Climate Action Act	01/01/2015	Legislative	Mitig...	Adaptation...
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...

Rows

Tables

Country	Name	Year Passed	Executive	Fra...	Categories	Document.
Japan	Act on the Improvem...	01/01/2015	Legislative	null	Energy De...	Law
Macedoni...	Action Plan on Rene...	01/01/2015	Executive	null	Energy Sup...	Plan
Czech Rep...	Adaptation strategy ...	01/01/2015	Executive	Adapt...	Adaptation	Strategy
Niue	Agriculture Sector Pl...	01/01/2015	Executive	null	Adaptation...	Plan
Thailand	Alternative Energy D...	01/01/2015	Executive	null	Energy Sup...	Plan
United Sta...	Clean Power Plan	01/01/2015	Executive	null	Energy Sup...	Plan
Malta	Climate Action Act	01/01/2015	Legislative	Mitig...	Adaptation...	Law
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...	Law
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...	Law



.XLS

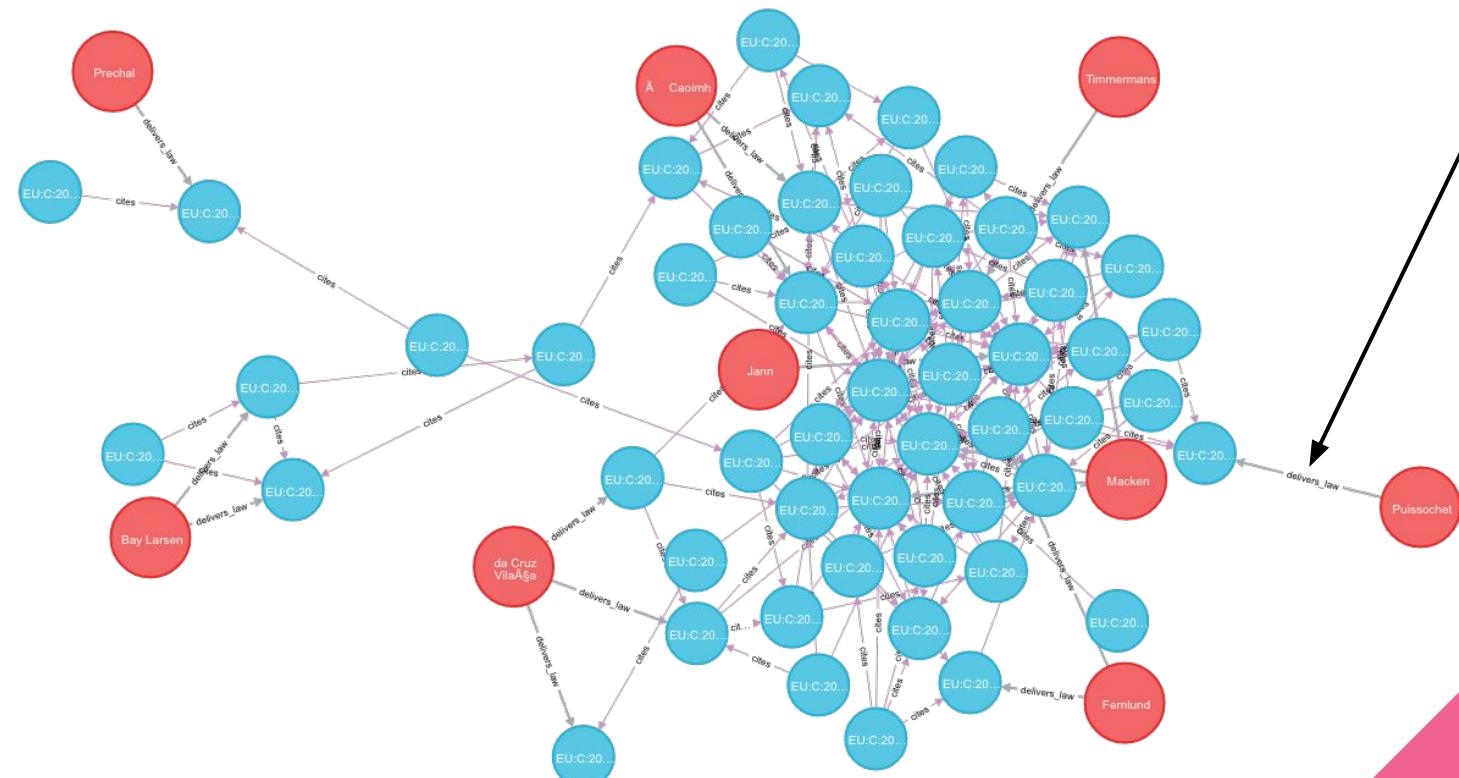
.CSV

.TSV

.data

Advantages:
File sizes and compatibility with other systems

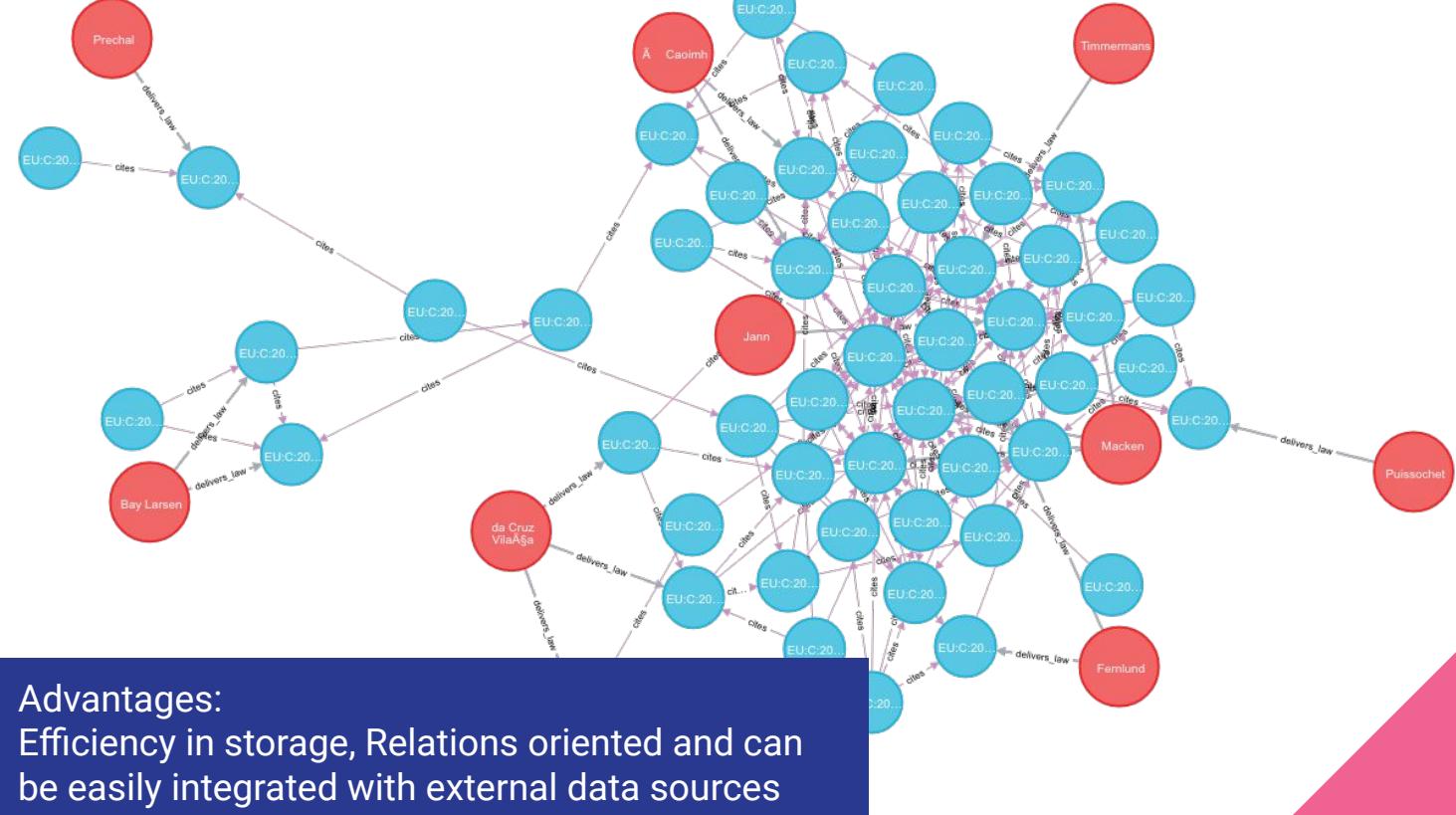
Graphs



Relations

Nodes

Graphs



.nt
.ttl
.gexf
.grapML

Advantages:
Efficiency in storage, Relations oriented and can
be easily integrated with external data sources

Documents

```
first_name: 'Paul',
surname: 'Miller',
cell: 447557505611,
city: 'London',
location: [45.123,47.232],
Profession: ['banking', 'finance', 'trader'],
cars: [
    { model: 'Bentley',
      year: 1973,
      value: 100000, ... },
    { model: 'Rolls Royce',
      year: 1965,
      value: 330000, ... }
]
```

Fields

String

Number

Geo-Coordinates

Typed field values

Fields can contain arrays

Fields can contain an array of sub-documents

Documents

```
first_name: 'Paul',
surname: 'Miller',
cell: 447557505611,
city: 'London',
location: [45.123,47.232],
Profession: ['banking', 'finance',
cars: [
    { model: 'Bentley',
      year: 1973,
      value: 100000, ... },
    { model: 'Rolls Royce',
      year: 1990,
      value: 200000, ... }
  ]]
```



.json

.xml

.yaml

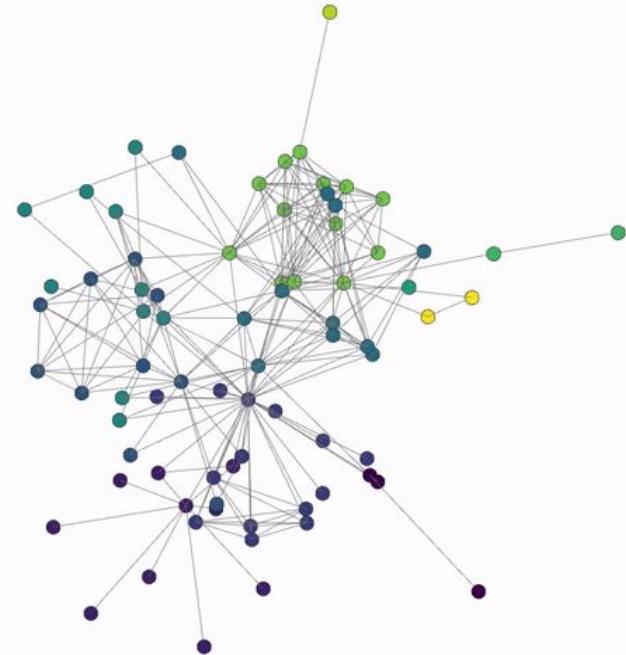
Advantages:

Structured dictionary object. Collection of complex documents with arbitrarily nested data

Knowledge Databases



GRAPHS ARE COMING



Knowledge Databases = Knowledge Graphs

Google knowledge graph

All Images News Videos Books More Settings Tools

About 314.000.000 results (0,59 seconds)

The **Knowledge Graph** is a **knowledge base** used by Google and its services to enhance its search engine's results with information gathered from a variety of sources. The information is presented to users in an infobox next to the search results.

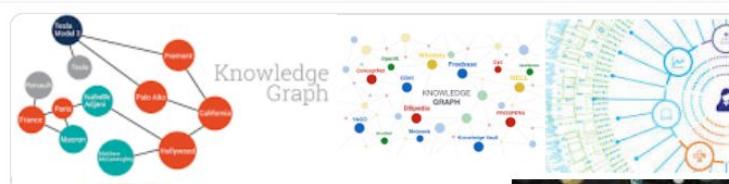
[Knowledge Graph - Wikipedia](#)
https://en.wikipedia.org/wiki/Knowledge_Graph

People also ask

- What is Knowledge Graph in SEO?
- Why is knowledge graph important?
- What is knowledge graph embedding?
- What is an Enterprise Knowledge Graph?

Feedback

Knowledge Graph



More images

The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources. The information is presented to users in an infobox next to the search results. [Wikipedia](#)

Feedback

Knowledge Graphs



Netherlands

Country in Europe

The Netherlands, a country in northwestern Europe, is known for a flat landscape of canals, tulip fields, windmills and cycling routes. Amsterdam, the capital, is home to the Rijksmuseum, Van Gogh Museum and the house where Jewish diarist Anne Frank hid during WWII. Canalside mansions and a trove of works from artists including Rembrandt and Vermeer remain from the city's 17th-century "Golden Age."

Capital: [Amsterdam](#)
Dialing code: +31
Official regional languages: West Frisian; Papiamento; English;
Points of interest: [Keukenhof](#), [Anne Frank House](#), [Rijksmuseum](#), [MORE](#)
Did you know: Netherlands has the fifth-largest natural gas exports (53,650,000,000 cu m) in the world. [wikipedia.org](#)

People also search for

View 15+ more

Belgium Europe Germany Holland Curaçao

[Feedback](#)

[Visit the Netherlands: Destinations, tips and inspiration - Holland.com](#)

<https://www.holland.com/global/tourism.htm> ▾

Welcome to Holland.com, the official website of the **Netherlands** Board of Tourism and Conventions, where you receive all the information for your visit to ...

20 Oct - 28 Oct [Dutch Design Week](#) Eindhoven

27 Oct - 28 Oct [Störrig Festival](#)

31 Oct - 4 Nov [Affordable Art Fair](#)

[Netherlands - Wikitravel](#)

<https://wikitravel.org/en/Netherlands> ▾

The **Netherlands** (Dutch: Nederland, also commonly, but incorrectly, called Holland) is a European country, bordering Germany to the east, Belgium to the south, ...

Country code: +31

Population: 16,803,893 (2013 estimate)

Area: 41,543km²; water: 7,650km²; land: 33,89...

Capital: Amsterdam

A knowledge graph is a special kind of **database** which **stores knowledge** in a **machine-readable** form and provides a means for information to be collected, organised, shared, searched and utilised.



Knowledge Graphs

- 1956, Rene Bakker (Utrecht) developed the notion of a Knowledge Graph in his PhD as a structure and representation of scientific knowledge.
- Late 1990s endeavours to index web pages using graph analytics techniques.
- Mid-late 2000s, commercial atomicity, consistency, isolation, and durability (ACID) graph databases such as Neo4j and Oracle Spatial and Graph became available.

MacArthur 'Genius Grant' Winner Maria Chudnovsky on Graph Theory

Wednesday, October 03, 2012



Bright Launches Bright Packed With Jobs Data Seeking Tips



A InfoWorld Home / InfoWorld Tech Watch / Buzz grows around graph databases



The First Word on Tech
INFOWORLD TECH WATCH

AUGUST 29, 2012

Buzz grows around graph databases

Interest in graph databases will continue to grow, given its ability to analyze data delivered in a non-relational format, such as social networking data

By Paul Krill | InfoWorld

Follow @pjkrill

Graph Databases: The New Way to Access Super Fast Social Data



259

Like

969

Tweet

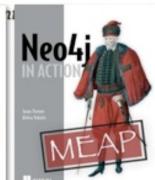
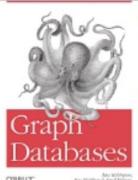
48

Share

173

Facebook's Social Graph, Neo4j show rising use of graph databases

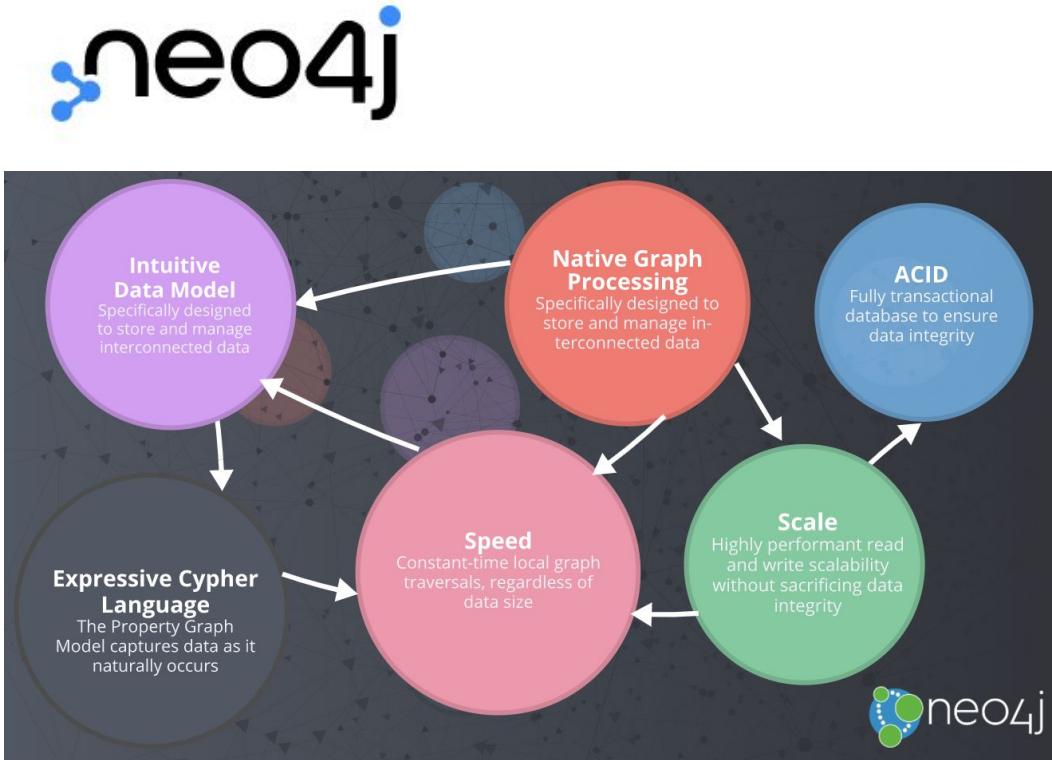
Summary: Facebook's Social Graph -- the database underlying its Graph Search engine unveiled yesterday -- is just one of many graph databases being employed for complex, connected data. Neo4j



I saw my own Interest Graph and it's scary accurate. We'd certainly pay for the ability to use the Gravity personalization technology I saw today at TechCrunch to help target content to users.

TC Michael Arrington, TechCrunch

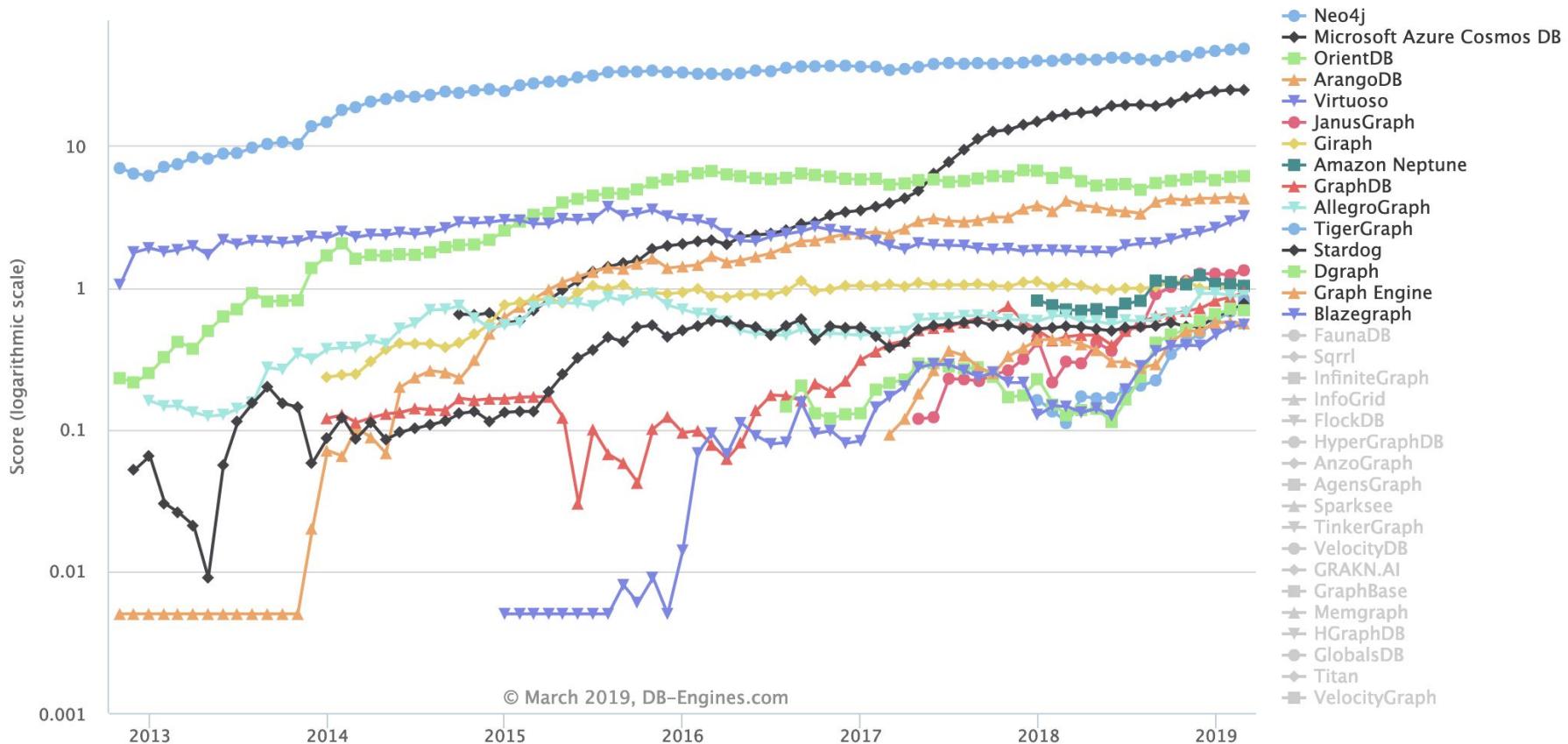
Knowledge Graphs



- Provides a compliant transactional backend for applications.
- Publicly available since 2007. Source code, written in Java and Scala
- The most popular graph-based database
- Cypher query language
- Largest ecosystem

Knowledge Graphs

https://db-engines.com/en/ranking_trend/graph+dbms

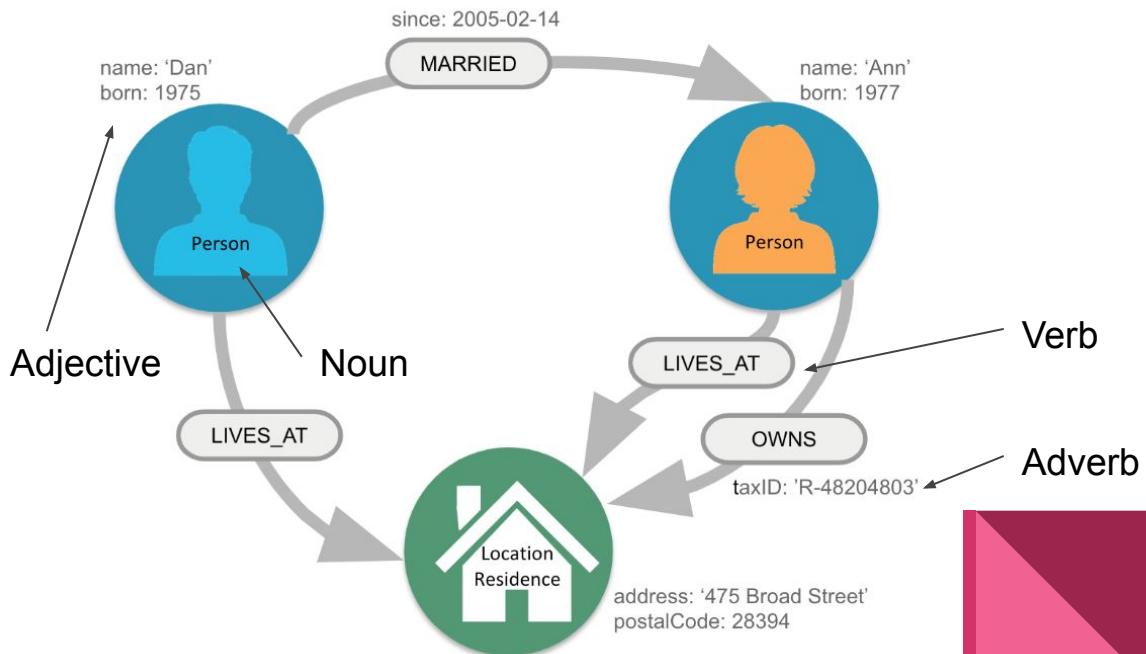


Knowledge Graphs

The graph model should be interpretable as a natural language

- **Noun:** Node label
- **Adjective:** Node property
- **Verb:** Relation label
- **Adverb:** Relation property

Person named 'Dan' born in 1975
is MARRIED since 2005-02-14 to
Person named 'Ann' born in 1977



Graph Query Language (CYPHER)

() Node

[] Relation

{ } Property

: Label

-> Direction

// Comment

Cypher

(:Person {name: 'Dan'}) - [:MARRIED {since: 2005-02-14}] -> (:Person {name: Ann})

Noun

Verb

Adverb

Adjective

Person named 'Dan' is MARRIED since 2005-02-14 to Person named 'Ann'

English

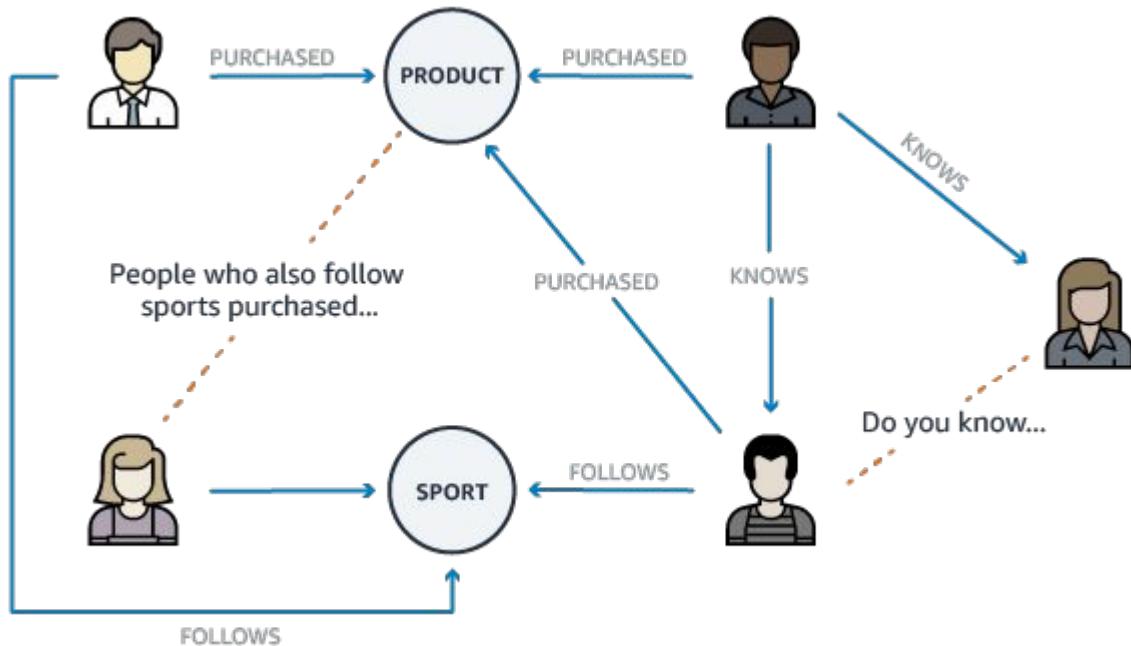
neo4j.com/sandbox

Get started with Neo4j Sandbox
while your coffee is still brewing

Neo4j is a native graph database, purpose-built to leverage data relationships and enable richer, more intelligent applications

[Launch the Free Sandbox](#)

Applications in Tech



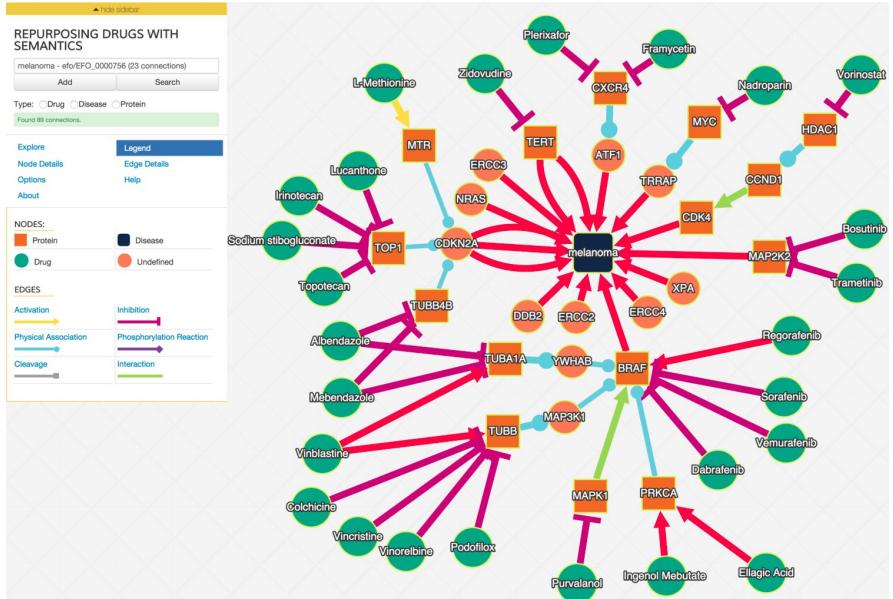
NETFLIX

amazon.com

Applications in Science

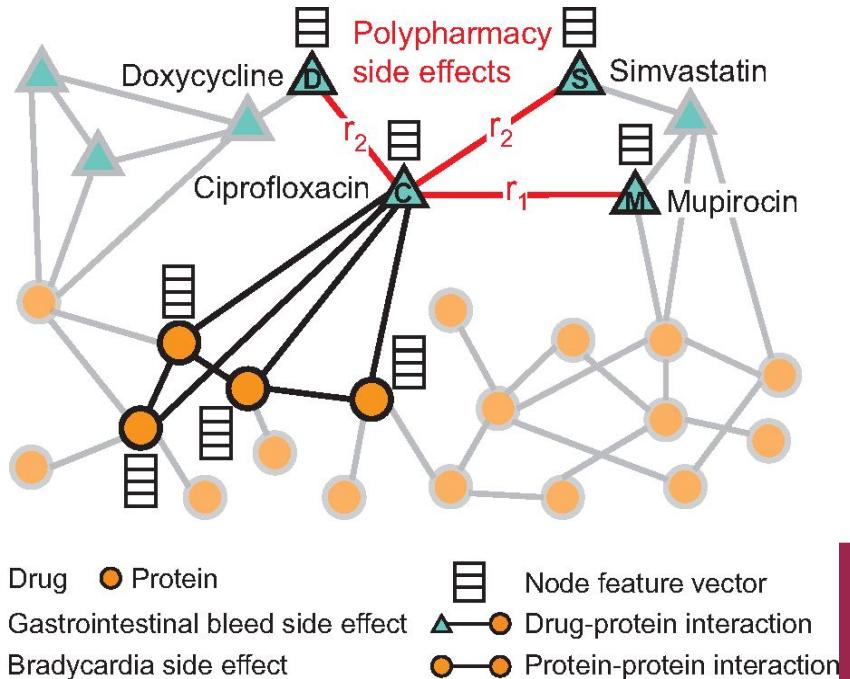
Modeling polypharmacy side effects with graph convolutional networks. Zitnik, Marinka, et al. *Bioinformatics* 34.13 (2018): i457-i466.

Drug discovery



Finding melanoma drugs through a probabilistic knowledge graph. *PeerJ Computer Science*. 2017. 3:e106 <https://doi.org/10.7717/peerj-cs.106>

Side effects



What are you making?

[All](#)[Biomass](#)[Ceramic](#)[Chemical](#)[Cosmetic](#)[Feed](#)[Food](#)[Metal](#)[Mineral](#)[Pharmaceutical](#)

Active pharmaceutical



Adhesives



Agar-agar



Alfalfa meal



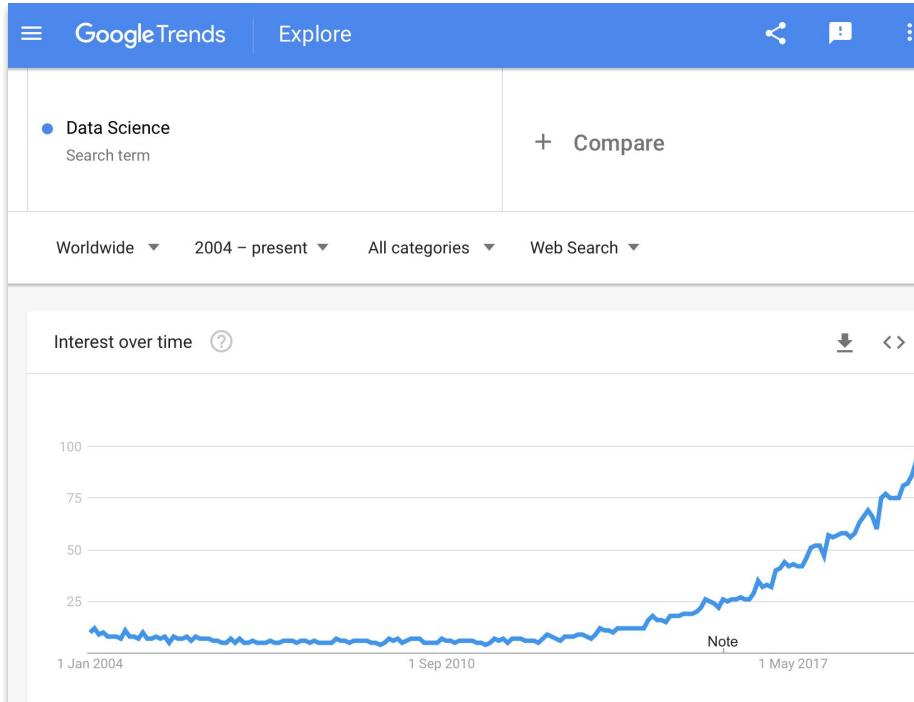
Almonds



What other applications
can we think of in
knowledge graphs?

Data Science with Python

Is Python for Data Science a Trend?



Related topics	
	Rising
1 Python - Programming language	Breakout
2 Analytics - Topic	Breakout
3 Big data - Topic	Breakout
4 Business - Organization type	Breakout
5 Salary - Topic	Breakout

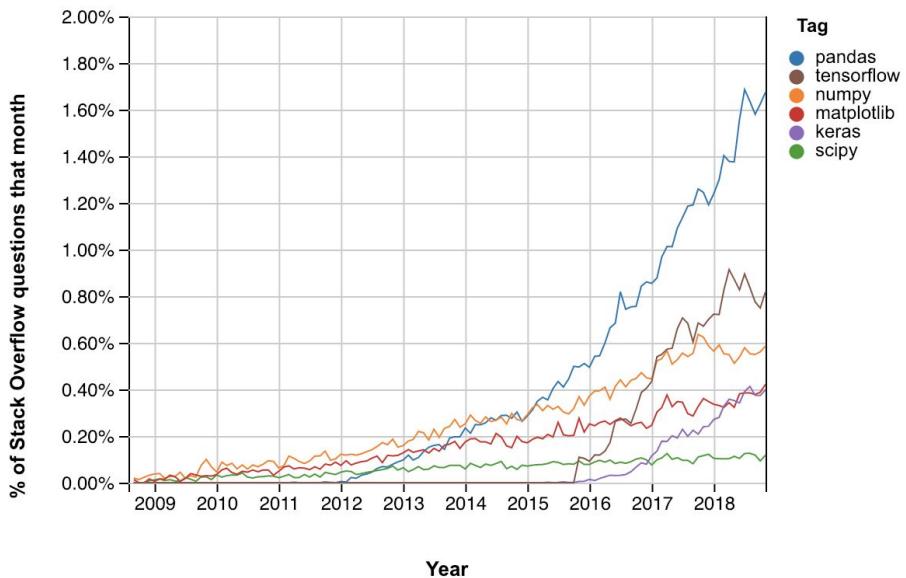
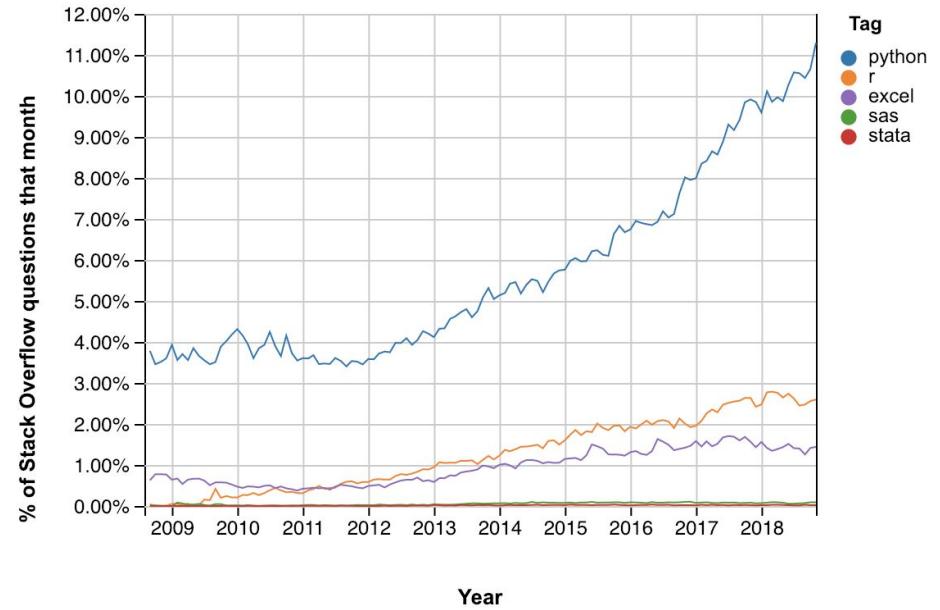
- Data is grows fast. It grows exponentially and continues to grow so. And it's estimated to be about 2.5 Exabytes, that is 2.5 million TB, a day. [1]
- And over 1.5 trillion queries on Google in a year.[2]

[1] <https://ourworldindata.org>

[2] <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/>

Source: <https://insights.stackoverflow.com/trends/>

Data Science with Python



Source: <https://insights.stackoverflow.com/trends>!

Data Science with Python

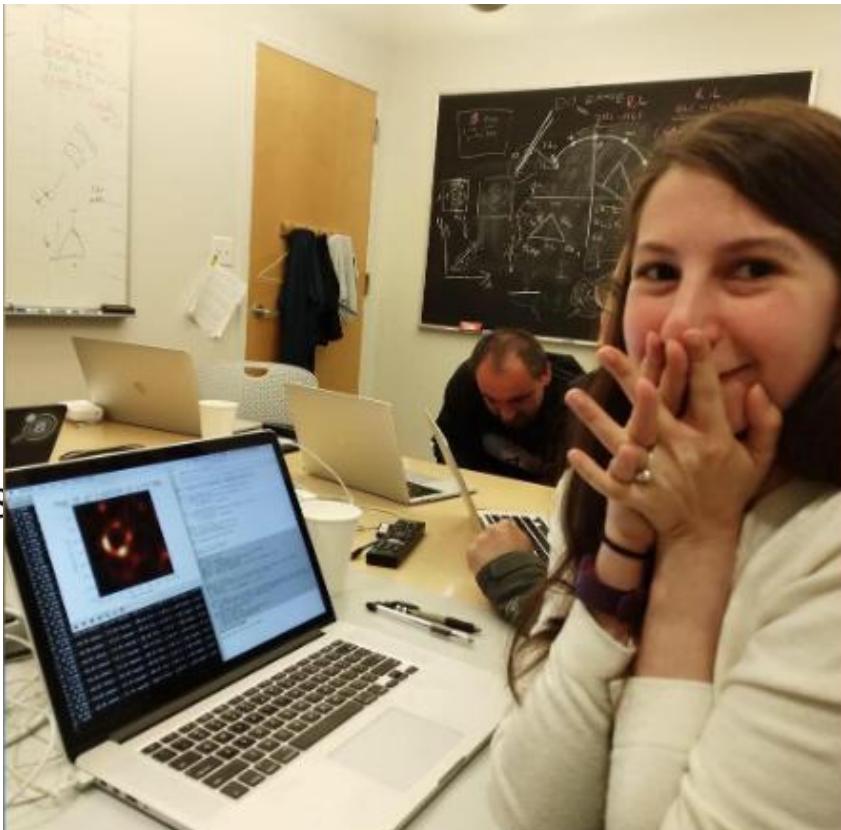
As a general purpose language, Python supports a large range of tasks.

Or put another way: 'Python isn't the best at anything, but it's second best at everything'

- Python is an, open-source high-level, interpreted, programming language.
- A data science/text analytics project may include everything from scraping data from the web, analyzing a mixture or text and numerical data, computing features, training a model, creating high-quality graphs, and then hosting a webapp with results.
- 90,000 libraries in the Python Package Index
- It has a massive user community, who contribute to a large number of high-quality, well maintained open-source tools.
- Widely used in industry.

Python brings “sub-technologies”

- **Numpy** (van der Walt et al. 2011)
- **Scipy** (Jones et al. 2001)
- **Pandas** (McKinney 2010)
- **Jupyter** (Kluyver et al. 2016)
- **Matplotlib** (Hunter 2007).
- **Astropy** (The Astropy Collaboration et al. 2013, 2018)



Python users

- Google Translate
- Spotify recommender
- Dropbox
- Netflix
- All NLP
- Video games (Sims)
- Visualization software (Dash)

It all got started, I believe, because the very earliest Googlers (Sergey, Larry, Craig, ...) made a good engineering decision: “Python where we can, C++ where we must.”

THE ASTROPHYSICAL JOURNAL LETTERS

OPEN ACCESS

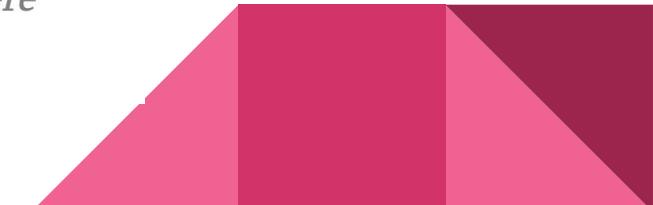
First M87 Event Horizon Telescope Results. III. Data Processing and Calibration

The Event Horizon Telescope Collaboration, Kazunori Akiyama^{1,2,3,4} , Antxon Alberdi⁵ , Walter Alef⁶, Keiichi Asada⁷, Rebecca Azulay^{8,9,6} , Anne-Kathrin Bacsko⁶ , David Ball¹⁰, Mislav Baloković^{4,11} , John Barrett²  +Show full author list

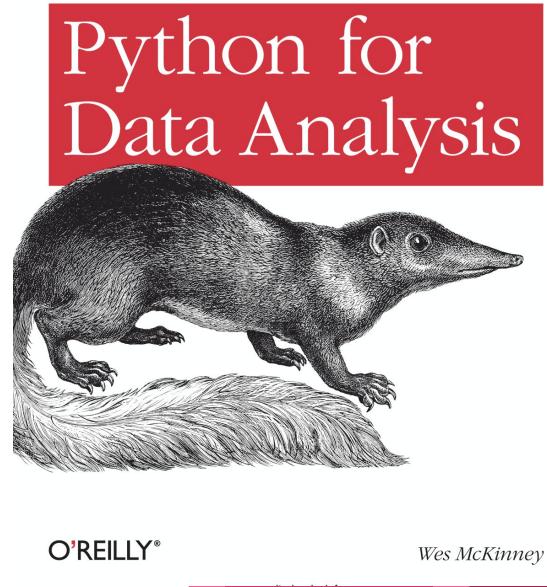
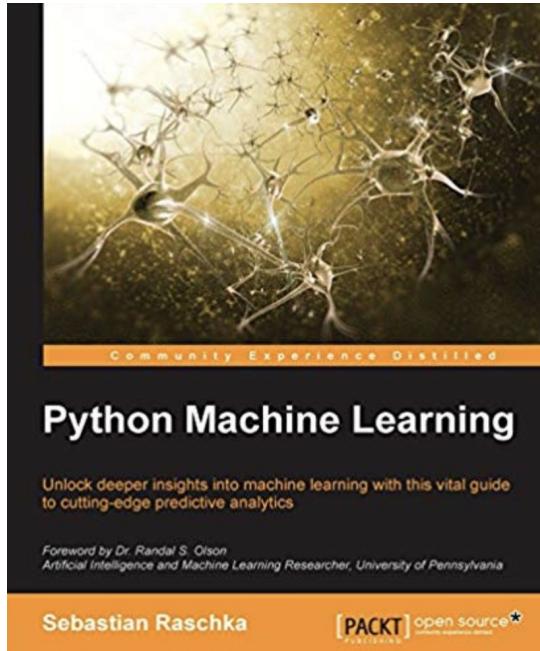
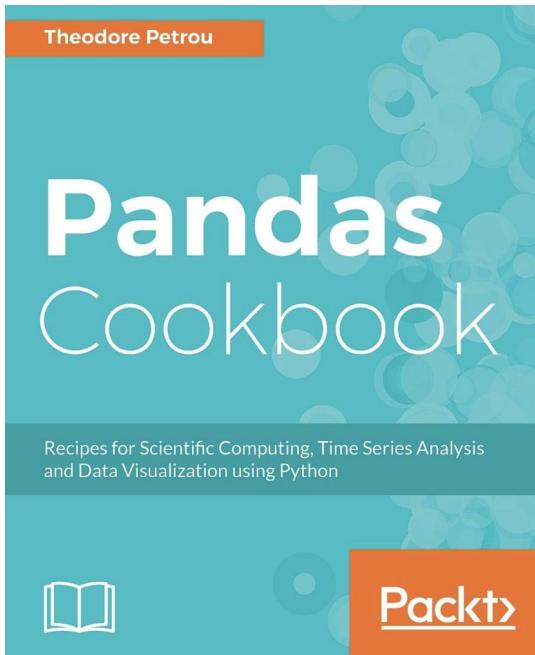
Published 2019 April 10 • © 2019. The American Astronomical Society.

[The Astrophysical Journal Letters, Volume 875, Number 1](#)

[Focus on the First Event Horizon Telescope Results](#)



Data Science Books



- Notebook 1: bit.ly/stamicarbon1
- Notebook 2: bit.ly/stamicarbon2
- Repo:
<https://github.com/pedrohserrano/data-science-technologies>
- Questionnaire:
forms.gle/yk5kr57r3UMjZTrL9